AD-A077 159 MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB F/G 6/16
DESIGN OF A ROBUST MAXIMUM LIKELIHOOD PITCH ESTIMATOR FOR SPEEC--ETC(U)
JUN 79 R J MCAULAY F19628-78-C-0002
UNCLASSIFIED TN-1979-28 ESD-TR-79-181 NL

1 OF 1
AD
A077159

END
DATE
FILMED
12-79
DDC

Technical Note 1979-28

R. J. McAulay

Design of a Robust Maximum Likelihood Pitch Estimator for Speech in Additive Noise

11 June 1979

Prepared for the Department of the Air Force
under Electronic Systems Division Contract F19628-78-C-0002 by

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

Approved for public release; distribution unlimited.

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the support of the Department of the Air Force under Contract F19628-78-C-0002.

This report may be reproduced to satisfy needs of U.S. Government agencies.

The views and conclusions contained in this document are those of the contractor and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Government.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Joseph C Syiek

Joseph C. Syiek
Project Officer
Lincoln Laboratory Project Office

12

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

DESIGN OF A ROBUST MAXIMUM LIKELIHOOD PITCH ESTIMATOR FOR SPEECH IN ADDITIVE NOISE

R. J. McAULAY

Group 24

TECHNICAL NOTE 1979-28

11 JUNE 1979

LEXINGTON MASSACHUSETTS

ABSTRACT

Using the maximum likelihood technique an algorithm is developed for the extraction of pitch for speech that has been corrupted by additive noise. The speech model includes the effects of pitch periodicity and the spectral envelope which results in a processing structure that consists of a noise suppression prefilter in cascade with a comb filter bank estimator-correlator. The prefilter attenuates those frequency bands where the speech signal-to-noise ratio is low, hence most of the deleterious noise is rejected prior to the determination of pitch by the comb filter bank correlator. The comb filter interpretation leads to an implementation of the correlation function which avoids the problem of anomalous pitch errors due to the effects of windowing and formant sidelobe interaction which obviates the need for any type of spectral flattening. Pitch ambiguities are resolved using a majority logic scoring algorithm and a carefully designed pitch tracker that can adapt rapidly to gross pitch variations. The voiced/unvoiced decision is based on an adaptive minimum energy threshold, a high/low band energy measurement, a normalized pitch correlation coefficient and a pitch track continuity coefficient. A time domain implementation of the algorithm that runs in real time in conjunction with an LPC analysis/synthesis system at 2400 bps is described.

CONTENTS

I. INTRODUCTION

With the development of high-speed minicomputer voice terminals it has become possible to deploy low bit-rate speech encoding algorithms in real-world operational environments. In some of these applications the speech is corrupted by an additive acoustical background noise which oftentimes results in a significant reduction in intelligibility [1]. This has stimulated research into the investigation of more robust algorithms for estimating the speech parameters. In this paper the focus is on the development of a robust pitch extractor based on the maximum likelihood technique. While this approach has already been explored by several authors [2] [3] [4], none of the models on which the analyses were based have taken into account the effects of the spectral envelope, and as a result pitch periodicity was the only discriminant that was used to combat the noise. When the envelope structure is included in the basic model, as is done in this paper, the resulting analysis leads to an algorithm that consists of a noise suppression prefilter in cascade with a comb filter bank correlator. The prefilter attempts to attenuate those frequency bands where the speech signal-to-noise ratio is low, hence most of the deleterious noise is rejected prior to the determination of pitch by the comb filter bank correlator. The comb filter interpretation leads to an implementation of the correlation function which avoids the problem of anomalous pitch errors due to the effects of windowing and the formant sidelobe interaction which obviates the need for any type of spectral flattening prior to pitch estimation.

Pitch ambiguities are resolved using a majority logic scoring algorithm and a carefully designed pitch tracker that can adapt rapidly to legitimate trailing edge pitch doubles. The voiced/unvoiced decision is based on an adaptive minimum energy threshold, a high/low band energy measurement, a normalized pitch correlation coefficient and a pitch track continuity coefficient. The paper describes a time domain implementation of the algorithm that runs in real time in conjunction with an LPC spectral analysis/synthesis system operating at 2.4 kbs or 3.6 kbs.

In Section II the pitch estimation problem is formulated within the framework of statistical estimation theory, and a sufficient statistic for the pitch estimator is derived. The ambiguity resolution logic and the design of the pitch tracker are presented in Section III, and the rationale for the buzz-hiss detector is discussed in Section IV. The algorithm is currently being subjected to extensive testing using the Diagnostic Rhyme Test (DRT) for clean speech and speech that has been corrupted by additive E4A Advanced Airborne Command Post (ABCP) noise.

II. DERIVATION OF THE SUFFICIENT STATISTIC

Based on a set of noisy observations of a voiced speech waveform it is desired to determine the "best" estimate of the pitch period. The maximum likelihood estimator is selected since it is easy to compute and for large signal-to-noise ratio (SNR), it is asymptotically unbiased and efficient (the variance converges to zero). The estimate is based on the data

$$y_n = v_n + w_n \tag{1}$$

where v_n and w_n represent the n'th sample of the voiced speech and noise waveforms respectively. To begin with the acoustic interference is assumed to be zero mean, white Gaussian noise with variance σ_w^2. Once the estimator has been derived for this case, the generalization to colored noise follows immediately using the analysis technique of the prewhitening filter. The voiced speech waveform is modelled as a sample function of a zero mean, Gaussian, quasi-periodic random process having covariance function $R_v(k) = R_v(k+\tau)$ where τ is the period of the process. This means that almost every sample function is periodic [5]. The likelihood function is the probability density function for the observations, which is

$$\ell(\tau) = p(y_1, y_2, \ldots, y_n|\tau) \tag{2}$$

Schweppe [6] has shown that for stationary Gaussian processes the log-likelihood ratio is

$$L(\tau) = -N\ell n\left(\sigma_p^2\right) - \frac{1}{\sigma_p^2} \sum_{n=1}^{N} (y_n - \hat{v}_{n|n-1})^2 \tag{3}$$

where $\hat{v}_{n|n-1}$ is the minimum mean squared error prediction of v_n based on measurements up to time n-1, namely $y_{n-1}, y_{n-2}, \ldots$ and σ_p^2 is the prediction error variance obtained by averaging over the ensemble of speech and noise sample functions. In practice the background noise level is unknown which renders σ_p^2 a nuisance parameter. The maximum likelihood estimate of σ_p^2 is found by maximizing (3) and is*

$$\hat{\sigma}_p^2(\tau) = \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{v}_{n|n-1})^2 \qquad (4)$$

from which the log-likelihood function reduces to

$$L(\tau) = -\ell n\left[\hat{\sigma}_p^2(\tau)\right] \qquad (5)$$

Therefore the maximum likelihood estimate of the pitch period (denoted as $\hat{\tau}$) can be found by choosing τ to minimize the energy in the prediction residual. In order to avoid the issue of explicitly implementing the predictor, it suffices to recognize that as a consequence of the Innovations Theorem [7] and the fact that speech is a Markov process, the residual sequence $\varepsilon_n = y_n - \hat{v}_{n|n-1}$ is zero mean white noise when $\hat{\tau}$ is "close to" the true pitch period. As a result, the transformation from the input sequence $\{y_n\}$ to the sequence of residuals $\{\varepsilon_n\}$ is a linear whitening filter. This provides a necessary condition which is used in the appendix to specify the structure of the maximum likelihood estimator. It is shown that the first stage of processing is a noise suppression prefilter. This is specified by the transfer function $P(\omega)$

*$\hat{v}_{n|n-1}$ depends implicitly on the pitch period τ.

which must satisfy

$$|P(\omega)|^2 = \frac{E(\omega)}{E(\omega)+\sigma_w^2} \tag{6}$$

where $E(\omega)$ is the spectral envelope of the voiced speech (the Fourier transform of the correlation function $R_v(\tau)$). The action of the filter is to suppress those frequencies where the SNR is low and to pass those frequencies where the SNR is high. Of course specification of $P(\omega)$ requires knowledge of the speech spectral envelope of $E(\omega)$ which is not available a priori. Techniques for estimating $E(\omega)$ from noisy data (which is not necessarily white) and for implementing $P(\omega)$ have been developed by McAulay and Malpass [8].

Having used the spectral envelope information to enhance the noisy speech waveform, the next stage of processing exploits the voiced speech periodicity to determine the pitch using correlation techniques. If $Y(\omega)$ represents the discrete Fourier Transform (DFT) of the prefilter input, and

$$X(\omega) = P(\omega)Y(\omega) \tag{7}$$

represents the DFT of the prefilter output, then it is also shown in the appendix that the pitch likelihood function reduces to

$$L(\tau) = \sum_{n=1}^{N} x_n \hat{x}_n(\tau) \tag{8}$$

where

$$\hat{x}_n(\tau) = \frac{1}{2}(x_{n-\tau}+x_{n+\tau}) \qquad (9)$$

represents the output of a comb filter tuned to pitch period τ. The action of this filter is to produce an estimate of the waveform x_n, presuming its periodicity to have period τ. A bank of comb filters is needed for each possible pitch candidate in the range of interest (i.e., 40 Hz - 300 Hz) and the pitch corresponding to the comb filter that leads to the largest correlation determines the maximum likelihood estimate of the pitch period. A block diagram illustrating the signal processing requirements to compute $\ell(\tau)$ is shown in Figure 1.

Substitution of (9) into (8) would show that, except for the prefilter, the maximum likelihood pitch estimator is basically another version of the correlation function pitch extractor [9], [10]. However, the standard implementation of this technique is to choose a processing interval that is wide enough to include two or three periods of the speech waveform. The autocorrelation function of the windowed data is then computed. The effect of this is to apply a triangular window to the true correlation function which causes the sidelobes introduced by the formant resonance to result in peaks that can be larger than the peak at the true pitch period, especially if the formant bandwidth is narrow. This, in turn, leads to large anomalous pitch errors. It has been the desire to remove the formant interaction that has led to the use of spectral flattening [9], [10] prior to the computation of the

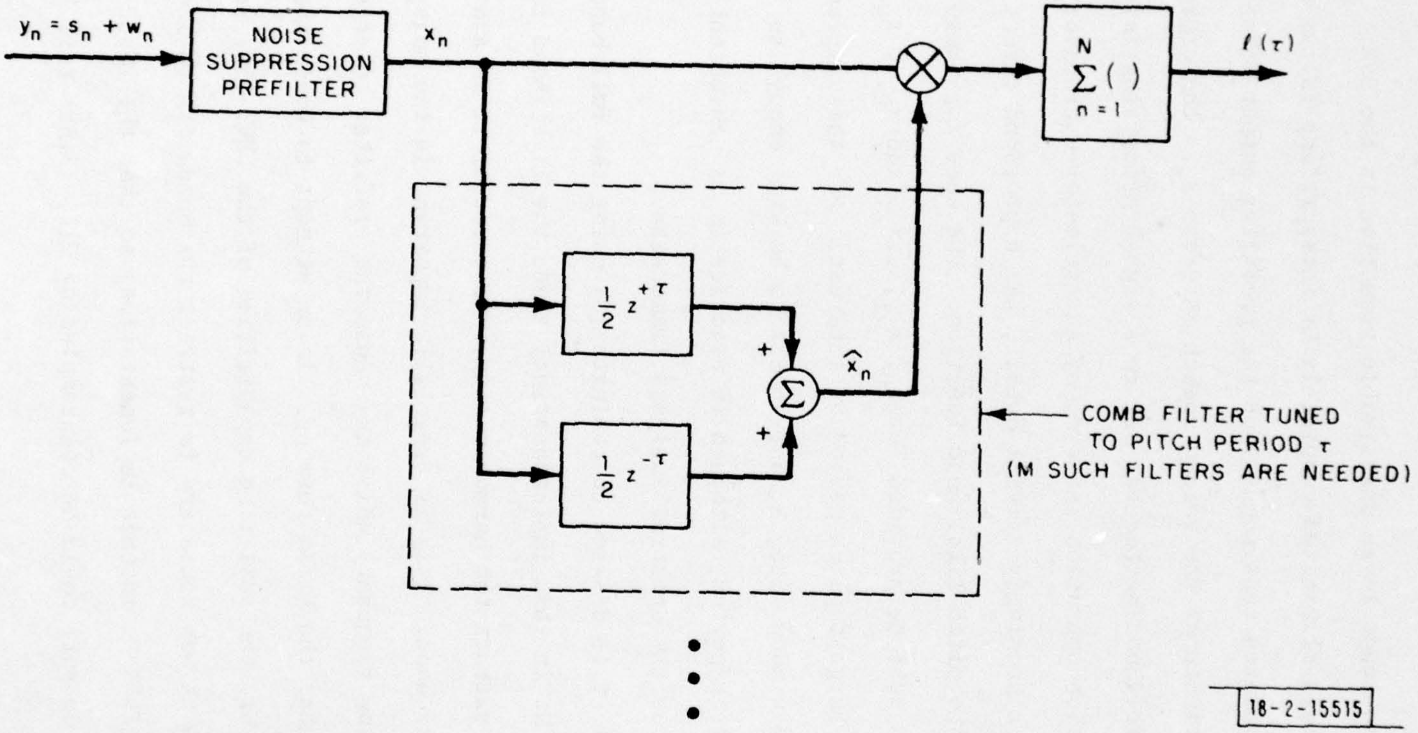

Fig.1. Structure of the maximum likelihood pitch estimator.

autocorrelation function. However, this not only complicates the processing but also leads to an undesirable reduction in the SNR. These problems have been eliminated completely by interpreting (9) as a comb filter which operates continuously on the prefilter output waveform to produce the estimate of the voiced speech waveform $\hat{x}_n$. Therefore use of the maximum likelihood method has led to a signal processing interpretation of the correlation operation in terms of an estimator-correlator structure, which although a seemingly trivial detail, has a profound effect on the properties of the pitch likelihood function. The cost for these benefits is buffering which must be provided for $x_0, x_{-1}, \dots x_{-T}$ and $x_{N+1}, x_{N+2}, \dots x_{N+T}$ where T is the largest pitch period of interest. For this implementation the correlation window size, N, only needs to be wide enough to include at least one pitch period, although in practice it is convenient to use the same width as the spectral analysis frame size.

In general it is desired to evaluate (8) using the full bandwidth speech (3787.5 Hz in this implementation) since the likelihood function can profitably extract the harmonic structure wherever it is available in the frequency band. This is especially important in the noisy speech problem where low frequency noise may cause the prefilter to attenuate the harmonics near the first formant. In an attempt to cover the pitch range 40 - 300 Hz, the real time capabilities of the LDVT [11] were exceeded, hence it was necessary to restrict the bandwidth to 1/4 the sampling rate (7575 Hz in this implementation) so that the input speech (the prefilter output) could be downsampled by 2:1. Since this bandwidth restriction also applied to the likelihood function it sufficed to evaluate (8) at every other pitch sample. The missing values were then obtained by using the 32:19:-3 2:1 up-sampling filter [12].

A typical example of the likelihood function that was computed subject to the preceding conditions is shown in Figure 2a for a male speaker for the vowel /u/ as in ch*ew*. Another example in Figure 2b computed for a female speaker for the vowel /æ/ as in th*a*t shows more clearly the effect of a narrow formant bandwidth. In spite of the multiplicity of peaks at the formant frequency, the peak corresponding to the true pitch period is evident. This has always been found to be the case for the large number of utterances that have been examined. It is obvious that if the autocorrelation function had been computed in the usual way, the triangular window would have totally obscured the peak at the true pitch. While these steady-state results could have been obtained by computing either $\Sigma x_n x_{n-\tau}$ or $\Sigma x_n x_{n+\tau}$, it was found that using $\Sigma x_n(x_{n-\tau}+x_{n+\tau})$, as required by (8) and (9), resulted in a more stable likelihood function at the leading and trailing edges of a phoneme and during pitch and vowel transitions.

Although the proposed implementation has eliminated the problem of formant interaction without using spectral flattening, it has resulted in the introduction of ambiguous peaks. In Figure 2a, for example, the peaks at 9.9 ms and 14.9 ms are as well-defined as the peak at the true pitch of 4.88 ms. Since minor perturbations in these peak values will occur, a simple peak picking algorithm will lead to pitch doubling and

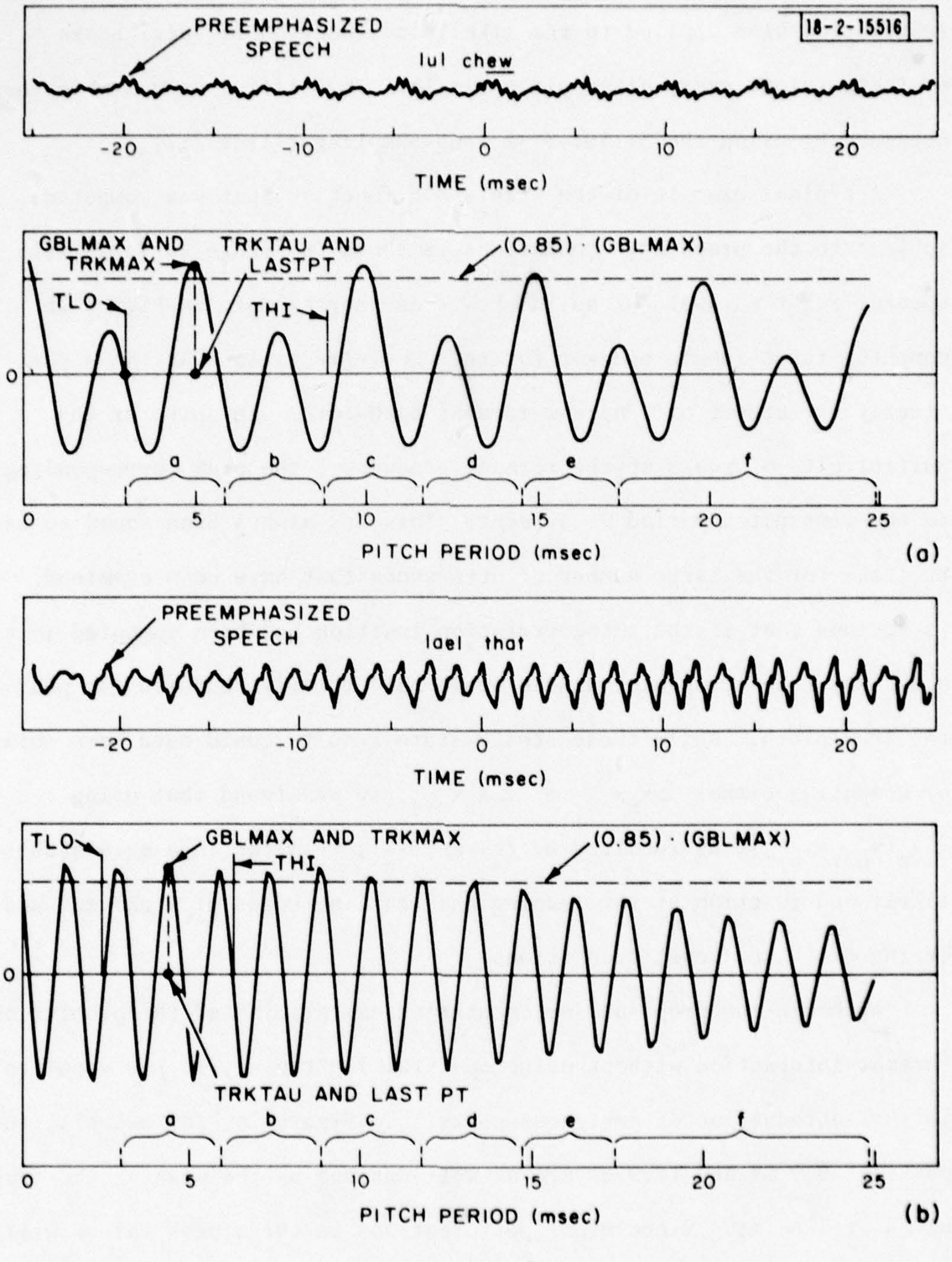

Fig.2(a&b). Typical voiced speech pitch likelihood functions.

pitch tripling errors. Since there is nothing that can be done to avoid this problem using only a single frame of data, heuristics must be introduced using data from multiple frames to resolve the ambiguous peaks. This issue is discussed in the next section.

III. PITCH AMBIGUITY RESOLUTION

The algorithm for resolving the pitch ambiguities is based on the majority logic decision scheme developed by Gold [13] [14]. To begin with a set of five elementary pitch estimators is constructed by searching the likelihood function for the largest peak lying in each of the unambiguous intervals 3-5.8 ms., 5.8-8.84 ms., 8.84-11.88 ms., 11.88-14.92 ms., and 14.92-25 ms. where a, b, c, d, e refer to the corresponding pitch periods at which each peak occurs. If an interval contains no peak (the slope must change sign for a peak to be defined), the pitch period is set equal to zero. A pitch tracker is presumed to exist (this will be described subsequently) which is specified in terms of a slow average pitch, PSLOW, a fast average pitch, PFAST, a lower tracker limit, TLO, and an upper tracker limit, THI. The largest peak that falls within the tracker window [TLO, THI] is labelled TRKMAX, while the pitch period at which this peak occurs is labelled TRKTAU. The largest of all of the peaks is labelled GBLMAX. If any of the five peaks lies below .85*GBLMAX then the corresponding pitch period is set to zero. A better feeling for these definitions can be obtained by referring to Figure 2 where all of the quantities have been labelled.

The next step is to associate two pitch period estimates with the output of each of the elementary pitch estimators: that for frame m, (the current frame) and for frame m-1 (the previous frame). These are labelled a, a', b, b', ... e,e'. Of the two only the most recent period is a candidate for the final pitch estimate (a,b,c,d or e) the other quantities are used only for scoring purposes. Figure 3 summarizes the scoring strategy as obtained for two frames of the vowel /u/ where Figure 2a represents the data for frame m. Each of the six current pitch candidates is compared against 25 quantities, itself included. Ten of the 25 are the two frame measurements from each of the five elemental pitch estimators. An additional 14 checks are obtained by computing the pitch values b/2, c/2, d/2, e/2, c/3, d/3, e/3, b'/2, c'/2, d'/2, e'/2, c'/3, d'/2, e'/3, which account for the presence of likelihood function peaks at double and triple the true pitch, as was the case in the examples in Figure 2. The final check is with respect to the fast average pitch, PFAST, as this takes into account the longer term properties of the pitch estimates computed over several frames. As indicated in Figure 3, a pitch candidate receives a vote if the test period against which it is compared is within a given percentage of the candidate value. The value of W (Figure 3) was chosen to be 1/8. The candidate that receives the highest score is taken as the trial pitch estimate for frame m and is labelled LASTPT. The corresponding value of the likelihood function is stored as LASTMX. Finally if the ambiguity resolved pitch estimate LASTPT lies within the tracker window [TLO,THI]

18-2-15517

		PITCH CANDIDATES				
		a	b	c	d	e
FRAME (m)	a	✓				
	b					
	c			✓		
	d					
	e					✓
	b/2					
	c/2	✓				
	d/2					
	e/2					
	c/3					
	d/3					
	e/3	✓				
FRAME (m-1)	a'	✓				
	b'					
	c'			✓		
	d'					
	e'					✓
	b'/2					
	c'/2	✓				
	d'/2					
	e'/2					
	c'/3					
	d'/3					
	e'/3	✓				
	PFAST	✓				
	SCORE	7	0	2	0	2

CHECK IN i^{th} ROW AND j^{th} COLUMN IF

$$\frac{P_i - P_j}{P_j} < W = \frac{1}{8}$$

Fig. 3. Scoring algorithm for pitch ambiguity resolution.

then the pitch estimate for frame m, P(m) is set equal to LASTPT, otherwise P(m) is taken to be the tracker pitch estimate TRKTAU. Usually the former step is taken, but sometimes, especially during vowel transitions the tracker estimate takes precedence thereby maintaining the continuity of the pitch track. To illustrate the concepts more clearly the algorithm when applied to the likelihood function in Figure 2a results in the following candidates for the pitch estimates: a=4.88, b=0, c=9.9, d=0, e=14.9. From Figure 3, candidate "a" with a score of 7 will be the choice for the unambiguous pitch estimate. Hence LASTPT=4.88 and since TLO=3.09 and THI=7.21, the tracker window encloses LASTPT and hence the pitch estimate for the m^{th} frame is P(m)=4.88 which, as is most often the case the same as TRKTAU. The reason for this is a result of the correctly placed tracker window. That this happens most of the time is due to the design of the pitch tracker. The tracker limits are set according to the rule

$$\mathrm{TLO}(m) = .6*\mathrm{PFAST} \tag{10a}$$

$$\mathrm{THI}(m) = 1.4*\mathrm{PFAST} \tag{10b}$$

where the "fast" and "slow" pitch averages evolve according to the recursions

$$\mathrm{PFAST}(m) = \mathrm{PFAST}(m-1) + \mathrm{QFAST}*[\mathrm{LASTPT} - \mathrm{PFAST}(m-1)] \tag{11a}$$

$$\mathrm{PSLOW}(m) = \mathrm{PSLOW}(m-1) + \mathrm{QSLOW}*[\mathrm{LASTPT} - \mathrm{PSLOW}(m-1)] \tag{11b}$$

where QFAST and QSLOW control the time constants of the "fast" and "slow" averaging filters. In the real time implementation frames were processed every 21 ms, hence PSLOW represented a long term estimate of the pitch for a particular speaker by setting QSLOW = .03 (.69 sec time constant). Should a new speaker having a radically different pitch use the same vocoder, then PSLOW is guaranteed to adapt to the new speaker since it tracks the unambiguous pitch estimate LASTPT which is not influenced in any way by the previous tracker settings. Although it is tempting to set up a tracker window about this long term average value, it has been found that for some speakers wide fluctuations in pitch can occur within a given utterance which demands more dynamic adaptivity in setting the tracker limits. It is for this reason that TLO and THI are tied to the fast average pitch PFAST which is up-dated using a much shorter time constant by setting QFAST = .35 (49 ms time constant). As a consequence PFAST represents an estimate of the short term average pitch, hence it can be used as a useful input to the scoring table for pitch ambiguity resolution as well as producing tracker limits that adapt more quickly to rapid pitch variations. An extreme, although not uncommon case, is illustrated in Figure 4 which shows the pitch track and the tracking parameters for the utterance "I heard that shot echo" spoken by a female. At the end of the vowel /o/ in ech_o_ the pitch doubles. However, after two frames the upper pitch tracking limit adapted to a value that was large enough to include the true pitch period.

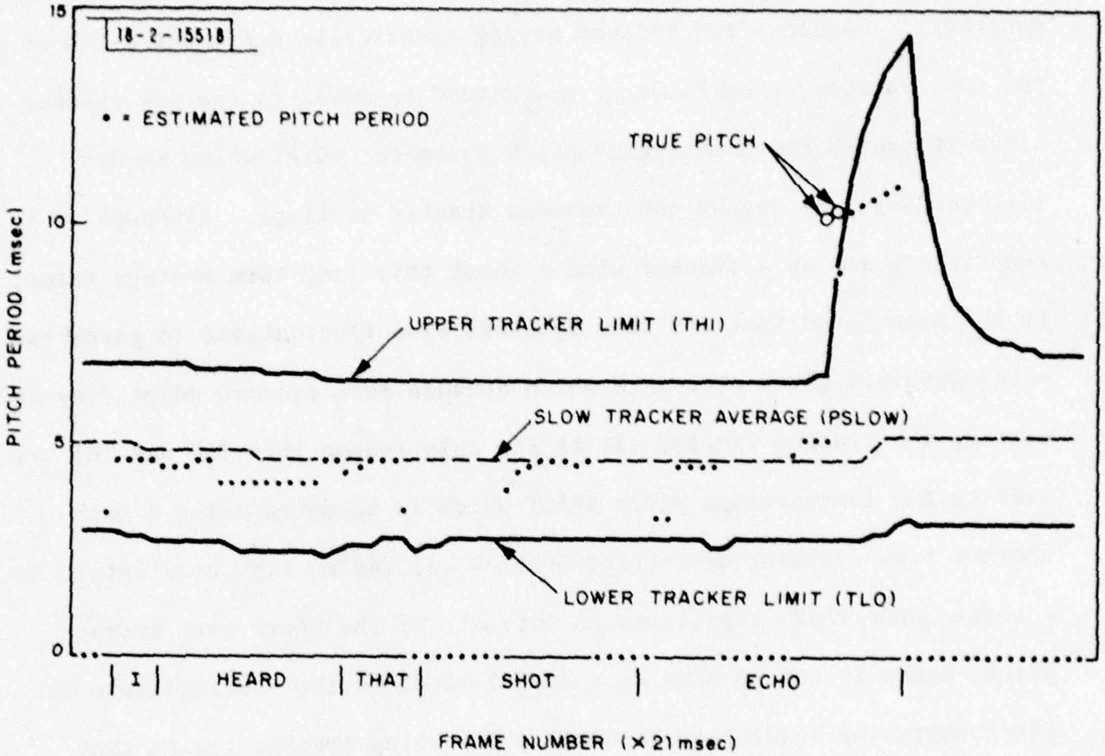

Fig.4. Typical trajectories for the pitch tracker parameters.

While the fast adaptation capabilities are needed to track the increasing pitch period at the trailing edge of an utterance, secondary problems can arise if the lower tracker limit adapts according to (10) without constraint. For example, if the female speaker who uttered the phrase "I heard that shot echo" had continued speaking, it is most likely that the pitch period would have returned to a value near the long term average of 4.62 ms. However, because of the trailing edge pitch double, the short term average pitch was 10.43 ms, and from (10) the lower tracker limit would have been 6.25 ms which would have blocked out subsequent pitch period estimates. Since PSLOW represents the long term average pitch, the lock-out problem can be avoided by requiring that

$$TLO \leq .6*PSLOW \qquad (12)$$

For the example PSLOW was 4.62 ms at the end of the utterance, hence TLO was constrained to be below 2.77 ms which completely eliminated the lock-out problem.

Since it is also possible for the pitch period to decrease rapidly during certain inflections (this seems to be a less frequent event) a similar lock-out problem exists with respect to the upper tracker limit. Hence the constraint

$$1.4*PSLOW \leq THI \qquad (13)$$

is used to eliminate the possibility that this lock-out event can occur.

As a final precaution that pitch inflections of the type described will not cause the pitch tracker to settle on an inappropriate orientation during unvoicing, the fast tracker parameters are relaxed to the steady-state values

$$\text{TLO} = .6*\text{PSLOW} \tag{14}$$

$$\text{PFAST} = \text{PSLOW} \tag{15}$$

$$\text{THI} = 1.4*\text{PSLOW} \tag{16}$$

To accomplish this, PFAST is adapted by driving recursion (11a) with PSLOW as the excitation and computing TLO and THI from (10) subject to constraints (12) and (13) whenever a frame of unvoiced speech is detected. The slow tracker average is not altered during such an unvoiced classification. The effect of the tracker relaxation is illustrated in Figure 4 during the silent interval at the end of the utterance.

In the preceding discussion use has been made of a buzz-hiss detector to determine when to update the tracker limits. The voicing detection algorithm will be discussed in the next section.

IV. VOICED-UNVOICED SPEECH DETECTION

The buzz-hiss detector is probably the most critical component of a narrowband vocoding system since it not only impacts significantly on intelligibility but has a profound effect on user acceptability. While buzz-hiss algorithms have been developed that work well on clean speech, problems often arise when the same algorithms are applied to speech which has been distorted due to additive noise. This is expecially true for noise sources having a well-defined spectral characteristic at low frequencies, such as E4A advanced airborne command post noise, since the noise waveform has many of the attributes of voiced speech. Since the maximum likelihood pitch estimator uses a noise suppression prefilter to enhance the signal-to-noise ratio prior to pitch correlation, it is reasonable to design the buzz-hiss logic to deal with speech contaminated by a low level residual noise. No single discriminant has been found that represents a necessary and sufficient condition for voiced speech, hence a number of tests were used in sequence. These tests were:

1. minimum energy threshold
2. high/low band energy measurements
3. pitch correlation coefficient
4. pitch track continuity.

A flow chart for the voicing logic is shown in Figure 5 and will now be described in detail.

Test 1: In the first test the energy for one frame of prefiltered speech is computed in block floating point format for m'th frame as

$$e(m) = \left(\sum_{n=1}^{N} v_n^{\,2} \right) 2^{SCLFCT} \tag{17}$$

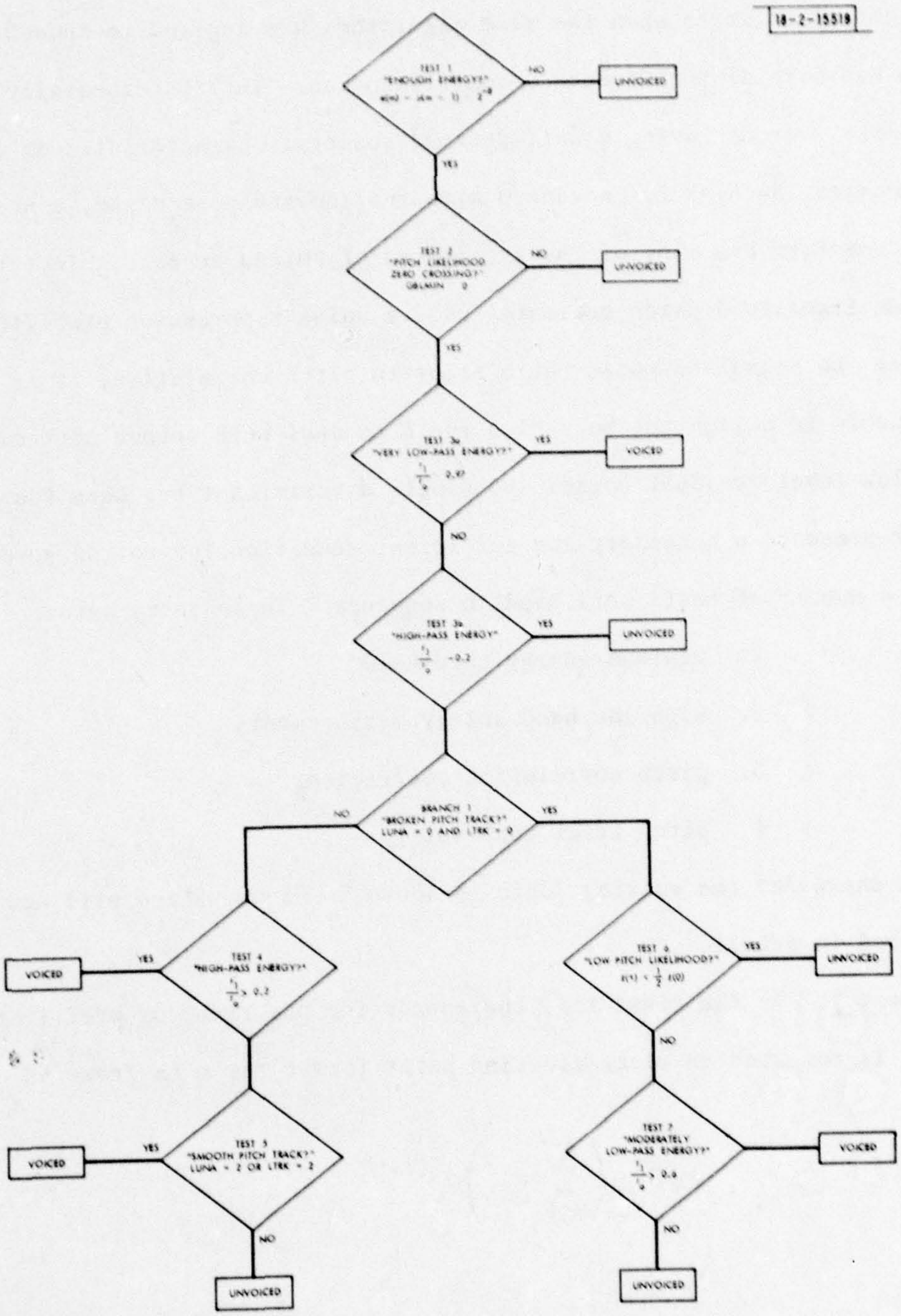

Fig.5. Logic for voicing detection.

where v_n is the prefilter output* and SCLFCT = 1,2,... if an overflow occurs. If $\mu(m-1)$ represents an estimate of the background noise energy computed on frame m-1, then an unvoiced or silence speech classification is made if SCLFCT = 0 and $e(m) - \mu(m-1) < 2^{-8}$ (v_n is treated as a 16 bit fraction). The description of the computation of $\mu(m-1)$ will be deferred until later in this section; however, for clean speech $\mu(m-1)$ will be zero.

Test 2: Once the energy test has been passed, a simple check is made on the structure of the correlation function. For all voiced sounds the correlation function has at least one zero crossing. While this is also the case for most unvoiced sounds, it sometimes happens that for low level noise and some unvoiced sounds such a zero crossing does not occur, and these cases are immediately classified as unvoiced speech. The test is implemented by determining if the minimum value of the pitch likelihood function (labelled GBLMIN) is positive.

Test 3: The next test measures the high/low spectral balance of the speech energy. Since energy in unvoiced sounds is generally located at high frequencies, measuring the spectral energy balance is a powerful voicing discriminant. If $L(\omega)$ represents a low pass filter and $H(\omega)$ a high pass filter, then

*In the real time program v_n is measured at the output of the 2:1 downsampling filter.

$$E_L = \int_0^{\pi} |L(\omega)|^2 |Y(\omega)|^2 d\omega \qquad (18)$$

$$E_H = \int_0^{\pi} |H(\omega)|^2 |Y(\omega)|^2 d\omega \qquad (19)$$

measure the low and high pass energies of the wideband speech signal y_n. The detection logic is to

$$\text{(a) declare voiced if } \frac{E_L}{E_H} > \lambda_v \qquad (20)$$

$$\text{(b) declare unvoiced if } \frac{E_L}{E_u} < \lambda_u \qquad (21)$$

The voicing decision is deferred whenever $\lambda_v < E_L/E_H < \lambda_u$. The threshold settings depend on the exact filters used in (18), but a particularly convenient choice is to take

$$|L(\omega)|^2 = \begin{cases} \cos\omega & 0 \leq \omega \leq \frac{\pi}{2} \\ 0 & \frac{\pi}{2} < \omega < \pi \end{cases} \qquad (22)$$

$$|H(\omega)|^2 = \begin{cases} 0 & 0 \leq \omega \leq \frac{\pi}{2} \\ |\cos\omega| & \frac{\pi}{2} < \omega < \pi \end{cases} \qquad (23)$$

As a consequence of these definitions, the difference between the low and high energies is given by

$$E_L - E_H = \int_0^{\pi} |Y(\omega)|^2 \cos \omega d\omega$$

$$= \sum_{n=1}^{N} y_n y_{n-1} \triangleq r_1 \qquad (24)$$

which is the measured correlation function at unit delay. Since the total energy in the band is

$$r_o \triangleq \sum_{n=1}^{N} y_n^2 \qquad (25)$$

which is the measured correlation function at zero delay, and since the total energy is approximately equal to the sum of the energies out of the low and high pass filters, then

$$r_o \approx E_L + E_H \qquad (26)$$

Combining (24), (26) and (20) results in the detection logic:

(a) declare voiced if $\frac{r_1}{r_o} > \frac{\lambda_v - 1}{\lambda_{v+1}} \triangleq \gamma_v$ (27a)

(b) declare unvoiced if $\frac{r_1}{r_o} < \frac{\lambda_{u-1}}{\lambda_{u+1}} \triangleq \gamma_u$ (27b)

which is a well-known buzz-hiss discriminant. However, the present derivation relates the r_1/r_o thresholds to the thresholds for the low and high pass energies. For speech that has not been preemphasized*, it has been found that a sufficient condition for unvoiced speech classification is achieved with $\lambda_u = -.2$ which corresponds to the requirement that the high pass energy be 50% larger than the low pass energy (i.e., $\lambda_u = 2/3$). For voiced speech it has been found necessary to require that $\lambda_v = .97$ which corresponds to the condition that the low pass energy be 66 times the high pass energy (i.e., $\lambda_v = 66$). While this seems overly conservative, any smaller value has been found to lead to many erroneous unvoiced-to-voiced classifications particularly for the plosives. Therefore the test in (27a) is mainly used to obtain correct voiced speech classifications of the nasals.

Branch 1: Since the high/low energy is not in itself a complete test for voicing, a further subdivision in the class of speech events is obtained by measuring the continuity of the pitch track. Two pitch continuity coefficients, LUNA and LTRK, are computed, one for the unambiguous pitch estimate LASTPT and one for the tracker pitch estimate

*In the real time implementation the speech undergoes analog preemphasis in order to enhance the dynamic range of the 12 bit A/D converter and to precondition the speech for LPC spectral analysis. Therefore a digital deemphasis filter is used prior to the computation of r_o and r_1.

TRKTAU. Letting PCAND denote the candidate pitch and letting TEST denote the pitch period against which it is being tested for coincidence, then a pitch correlation occurs if

$$|\mathrm{PCAND} - \mathrm{TEST}| < \frac{\mathrm{PCAND}}{8} \qquad (28)$$

This definition is applied to the evaluation of the unambiguous pitch continuity coefficient as follows: If LASTPT(m) does not correlate with LASTPT(m-1) then LUNA = Ø. If LASTPT(m) correlates with LASTPT(m-1) then LUNA = 1. If LUNA = 1 and LASTPT(m) correlates with LASTPT(m-2) then LUNA = 2. The tracker pitch continuity coefficient, LTRK, is computed in the same way replacing LASTPT(i), by TRKTAU(i), i = m, m-1, m-2. At the branch point the voicing algorithm declares a broken pitch track if LUNA = 0 and LTRK = 0, that is if neither of the frame m and frame m-1 pitch estimates correlate.

Test 4: The principal reason for the test for the broken pitch tract is to eliminate the plosives as possible candidates for voiced speech classification. A typical example of the likelihood function for a plosive is illustrated in Fig. 6 which shows that although the peaks are not insignificant, the largest value occurs at a randomly oriented pitch which tends to decorrelate with respect to previous values. Therefore if the branch declares in favor of an unbroken pitch track it is highly unlikely that the sound is a plosive. This means that the threshold on the spectral balance measure r_1/r_o can be relaxed and a

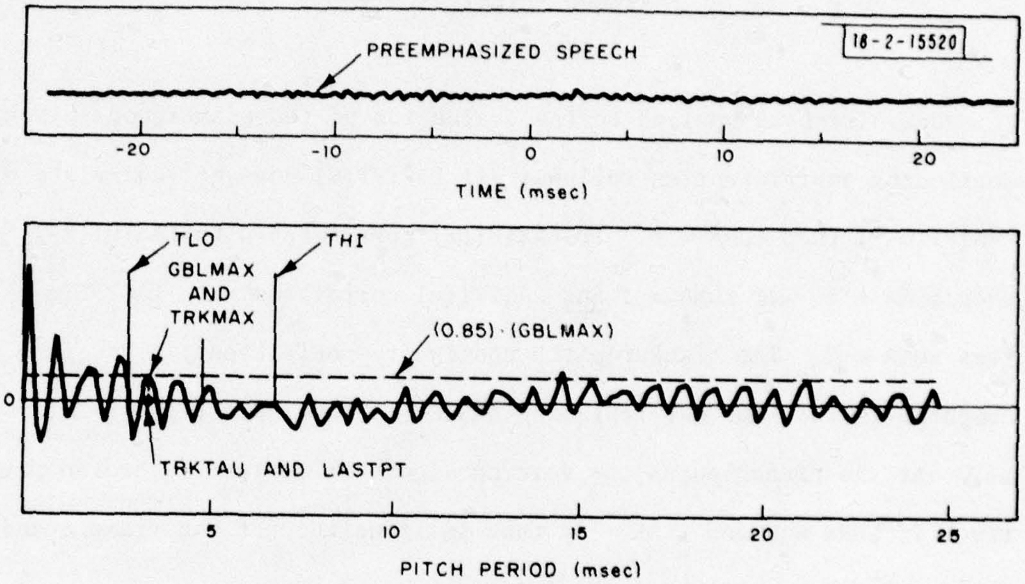

Fig.6. The pitch likelihood function for the plosive /p/ as in "prison."

weaker threshold is set at $\lambda_v = .2$ which results in a voiced speech classification whenever the low band energy is 50% larger than the high band energy.

Test 5: Since it may happen that microphone distortion or prefilter effects could affect the spectral balance in such a way that the preceding test for voiced speech might fail, as a final precaution another test for pitch continuity is made. In this case if either LUNA = 2 or LTRK = 2, that is if either the unambiguous pitch estimate or the tracker pitch estimate correlates with the corresponding pitch estimates on the previous two frames then the pitch track is declared "smooth" and a voiced speech classification is made.

Test 6: It was argued that the branching test for a broken pitch track was needed in the classification of the plosives. However a pitch track discontinuity can also occur if the pitch is in a rapid transition. When the latter event occurs the normalized pitch likelihood function is usually high, which is characteristic of voiced speech, but not of a plosive. Therefore if $\ell(\tau)$ represents the computed likelihood function at pitch period τ, then its value at the unambiguous pitch was defined to be LASTMX = ℓ(LASTPT) and an unvoiced classification is made if

$$\frac{\text{LASTMX}}{\ell(0)} < \frac{1}{2} \tag{29}$$

This test has the desired effect of classifying most of the plosives as unvoiced speech.

Test 7: While it is tempting to use (29) as a necessary and sufficient condition for unvoiced speech, it has been found that there are many unvoiced sounds for which the normalized correlation coefficient exceeds the 50% threshold. In order to correctly classify those voiced sounds corresponding to a pitch in transition (causing a broken pitch track and a large normalized correlation coefficient) another test on the spectral energy balance is used. In this case if $r_1/r_o > .60$, which corresponds to the requirement that the low pass energy be four times larger than the high pass energy, then a voiced speech classification is made. Failing this test results in an unvoiced speech classification.

Background Noise Energy Measurement: In order to complete the specification of the classifier algorithm it is necessary to describe the method for computing the average background noise energy $\mu(m)$ since this quantity was used in the very first classifier test. Basically the averaging is done using the first order recursion

$$\mu(m) = \mu(m-1) + \alpha(m)[e(m) - \mu(m-1)] \tag{30}$$

where $\mu(m)$ is the average energy and $e(m)$ is the measured energy for the m'th frame computed according to (17). A meaningful estimate for the background noise can be obtained only if (30) is not updated during steady voicing. Otherwise the detection threshold would rise resulting in erroneous unvoiced speech classifications. One way of reducing this problem [15] is to choose $\alpha(m)$ adaptively to correspond to a 4-second time constant if the energy is increasing $\left(\text{i.e., if } e(m) \geq \mu(m-1)\right)$ and to a 40 ms. time constant if the energy is decreasing $\left(\text{i.e., if } e(m) < \mu(m-1)\right)$. In this way $\mu(m)$ adapts to the noise energy almost instantly whenever there is a voiced speech gap (which occurs about 50% of the time), while during connected speech the noise level increases slowly enough that the noise power does not take on the attributes of speech. (The growth rate is also clamped so that only a 25% increase is allowed in a 21 ms. frame.) Additional restrictions on when the average energy can be updated are given in the flow chart of Fig. 7.

Test 1: Since the speech data has been preprocessed by a noise suppression filter the residual noise level is small enough that the overflow bit in the energy computation (17) is never set. Therefore if an overflow occurs it must be due to the presence of speech, hence the computation of $\mu(m)$ is by-passed (i.e., $\alpha(m)$ is set to zero).

Test 2: Since all voiced speech sounds result in a likelihood function having at least one zero crossing, the lack of a zero crossing can only correspond to unvoiced speech or noise, hence the average energy is up-dated.

Test 3: Although there are instances where the normalized likelihood function falls below the 50% threshold for voiced sounds, "most of the time" such a condition corresponds to unvoiced speech or noise, and the average energy is up-dated.

Test 4: Since pitch continuity is a strong indication of voicing,

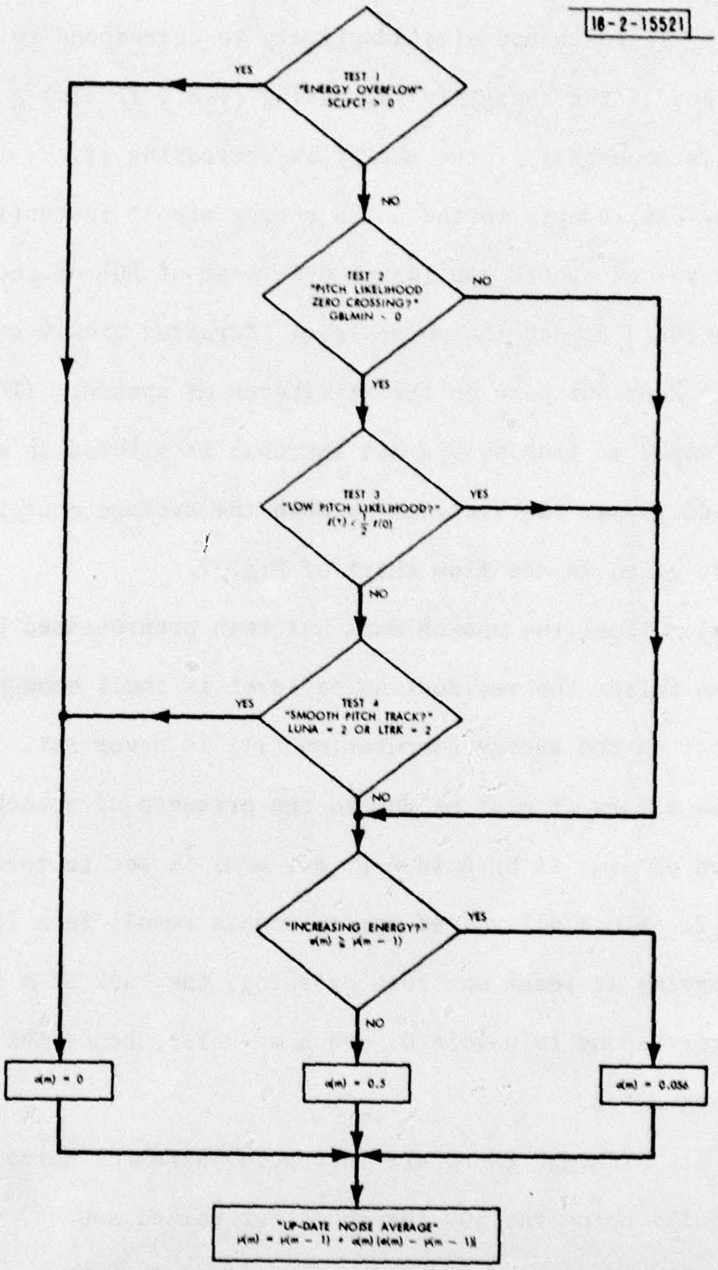

Fig.7. Logic for computing the average background noise level.

(43) is not updated whenever LUNA = 2 or LTRK = 2. Although this test may fail on occasion, it will not significantly affect the estimate of the background noise energy.

V. EXPERIMENTAL RESULTS AND CONCLUSIONS

The preceding classification algorithm has been evaluated using the real time program in conjunction with an LPC analysis/synthesis system operating at 2400 bps and 3600 bps. Literally hours of speech data have been processed encompassing male and female speakers, airborne command post and helicopter noise environments and dynamic and noise-cancelling microphones. Subjectively the algorithm seems to be quite robust. The major weakness is the tendency to classify plosives as voiced speech, an effect which is perceptible mainly for high quality, clean, female speech.

For speech corrupted by additive E4A Advanced Airborne Command Post noise, the buzz-hiss detector tends to classify the noise as voiced speech which results in an unpleasant buzzy quality in the speech synthesis. When the noisy data is processed by the noise suppression filter, with a small amount of suppression, the noise is classified as hiss and is more pleasant to the ear. As more suppression is introduced the noise can be made imperceptible but at the expense of buzz-hiss errors especially for the vowel /i/ as in eve. It is conjectured that this is due to the fact that the first formant lies in the region about 270 Hz. where there is also a large concentration of noise energy. The prefilter acts to suppress the noise at this frequency and in so doing also suppresses the speech signal (the amount of suppression depends on the SNR). This means that the pitch estimate and the voicing decision depend on the second formant waveform which lies in the vicinity of 2290 Hz. Since this is outside the range of the lowpass 2:1 downsampling filter there is little speech energy left to combat the residual noise. A potential solution to this problem is to compute the pitch estimate on the basis of the full bandwidth speech. Unfortunately this exceeds the computational capabilities of the LDVT when the LPC algorithms are implemented in a full duplex node. In an attempt to determine the value of such a wideband pitch estimator a half duplex version of the algorithm is currently under development.

ACKNOWLEDGMENT

During the course of this work, the author's colleague, Dr. D. B. Paul developed a real-time frequency domain implementation of the maximum likelihood pitch extractor. The multitude of technical discussions with Dr. Paul concerning the details of his implementation resulted in considerable improvements in the time domain algorithm.

REFERENCES

1. C. Teacher and H. Watkins, "ANDVT Microphone and Audio System Study," Ketron Final Report for the Commander, Naval Electronic Systems Command, Washington, D. C. (August 1978).

2. A. M. Noll, "The Cepstrum and Some Close Relatives," in J. W. R. Griffiths et al, Eds. Signal Processing (NATO Advanced Study Inst.) (New York, Academic, 1973).

3. J. D. Wise, J. R. Caprio and T. W. Parks, "Maximum Likelihood Pitch Estimation," IEEE Trans. Acoust., Speech and Signal Processing ASSP-24, 418 (1976).

4. D. H. Friedman, "Pseudo-Maximum-Likelihood Speech Pitch Extraction," IEEE Trans. Acoust., Speech, and Signal Processing ASSP-25, 213 (1977).

5. H. L. Van Trees, Detection, Estimation and Modulation Theory, Part I (Wiley, New York, 1968) p. 209.

6. F. Schweppe, "Evaluation of Likelihood Functions for Gaussian Signals," IEEE Trans. Inf. Theory IT-11, 61 (1965).

7. T. Kailath, "An Innovations Approach to Least-Squares Estimation - Part I: Linear Filtering in Additive White Noise," IEEE Trans. Automatic Control AC-13, 646 (1968).

8. R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Maximum-Likelihood Noise Suppression Filter," Technical Note 1979-31, Lincoln Laboratory, M.I.T. (19 June 1979).

9. L. R. Rabiner, "On the Use of the Autocorrelation Analysis for Pitch Detection," IEEE Trans. Acoust., Speech, and Signal Processing ASSP-25, 24 (1977).

10. J. D. Markel and A. H. Gray, Jr., Linear Prediction of Speech, (Springer-Verlag, New York, 1976), p. 199.

11. P. E. Blankenship, "LDVT: High Performance Minicomputer for Real-Time Speech Processing," EASCON '77 Record (26 September 1977).

12. D. J. Goodman and M. J. Carey, "Nine Digital Filters for Decimation and Interpolation," IEEE Trans. Acoust., Speech, and Signal Processing ASSP-25, 121 (1977).

13. B. Gold, "Description of a Computer Program for Pitch Detection," Paper G34, Proc. of the Fourth International Congress on Acoustics, Copenhagen, 1962.

14. T. Bially and W. M. Anderson, "A Digital Channel Vocoder," IEEE Trans. Communications Technology COM-18, 435 (1970).

15. D. Paul, "A Robust Vocoder with Pitch-Adaptive Spectral Envelope Estimation and an Integrated Maximum Likelihood Pitch Estimator," Proc. International Conference on Acoustics, Speech, and Signal Processing, Washington, D.C. (1979) p. 64.

APPENDIX
DERIVATION OF THE PITCH STATISTIC

Assuming that the pitch estimate $\hat{\tau}$ is "close to" the true pitch period then, as a consequence of the innovations theorem [7], the residual sequence $\varepsilon_n = y_n - \hat{v}_{n|n-1}$ is a white noise process. This means that the transformation from $\{y_n\}$ to $\{\varepsilon_n\}$ is a *linear whitening filter*. Letting the mapping be characterized by the transfer function $H(\omega;\hat{\tau})$ it must necessarily follow that*

$$|H(\omega;\hat{\tau})|^2 = \frac{\gamma}{S_y(\omega;\hat{\tau})} \tag{A-1}$$

where the constant γ represents the power spectrum of the residual sequence and

$$S_y(\omega;\hat{\tau}) = S_v(\omega;\tau) + \sigma_w^2 \tag{A-2}$$

represents the power spectrum of the ensemble of noisy voiced speech waveforms. Since the constant γ in (A-1) affects only the gain of the filter, then with no loss γ can be set equal to σ_w^2 which leads to the expression

$$|H(\omega;\hat{\tau})|^2 = 1 - \frac{S_v(\omega;\hat{\tau})}{S_v(\omega;\hat{\tau})+\sigma_w^2} \tag{A-3}$$

*To be precise, this should be written $H[\exp(j\omega)]$ where $\omega = 2\pi f/f_s$.

The corresponding minimum value for the prediction residual energy follows from Parseval's Theorem. It is

$$\sum_{n=1}^{N}\left(y_n-\hat{v}_{n|n-1}\right)^2 = \int_0^{\pi} |H(\omega;\hat{\tau})|^2 |Y(\omega)|^2 d\omega$$

$$= \int_0^{\pi} |Y(\omega)|^2 d\omega - \int_0^{\pi} \frac{S_V(\omega;\hat{\tau})}{S_V(\omega;\hat{\tau})+\sigma_W^2} |Y(\omega)|^2 d\omega \qquad \text{(A-4)}$$

where $Y(\omega)$ is the DFT of the measurement sequence. This result shows that if the voiced speech spectrum is completely known except for the pitch period τ, then the value of τ for which

$$\ell(\tau) = \int_0^{\pi} \frac{S_V(\omega;\tau)}{S_V(\omega;\tau)+\sigma_W^2} |Y(\omega)|^2 d\omega \qquad \text{(A-5)}$$

is a maximum is the maximum likelihood estimate. A good model for the voiced speech spectrum is

$$S_V(\omega;\tau) = E(\omega)C(\omega;\tau) \qquad \text{(A-6)}$$

where $E(\omega)$ represents the envelope of the speech spectrum (the DFT of the correlation function $R_V(\tau)$) which modulates the periodic line structure represented by $C(\omega;\tau)$. Maximizing (A-5) is therefore equivalent to maximizing

$$\ell(\tau) = \int_0^{\pi} |G(\omega;\tau)|^2 |Y(\omega)|^2 d\omega \qquad \text{(A-7)}$$

where $G(\omega;\tau)$ represents the linear filter which satisfies

$$|G(\omega;\tau)|^2 = \frac{E(\omega)C(\omega;\tau)}{E(\omega)C(\omega;\tau)+\sigma_w^2} \tag{A-8}$$

Therefore the likelihood function represents the energy at the output of a bank of filters, each being tuned to a different pitch period. The maximum likelihood estimate of τ corresponds to that filter in the bank for which the output energy is largest. Since the filters defined by (A-8) pass those frequencies at which the SNR is high and reject those at which it is low, then the effect of the comb filter in the denominator of (A-8) is to introduce nulls between the pitch harmonics which contribute to the definition of those frequencies at which signal rejection occurs. Since the nulls also appear in the numerator, approximately the same filtering performance can be obtained if the comb filter in the denominator is omitted. As a result (A-8) can be approximated by

$$|G(\omega;\tau)|^2 = C(\omega;\tau) \cdot \frac{E(\omega)}{E(\omega)+\sigma_w^2} \tag{A-9}$$

The second term in (A-9) can be interpreted as a filtering operation by defining

$$|P(\omega)|^2 = \frac{E(\omega)}{E(\omega)+\sigma_w^2} \tag{A-10}$$

Its function is to use the information in the spectral envelope to pass those frequencies for which the speech SNR is large and reject those for which it is small. Therefore $P(\omega)$ can be interpreted as a noise suppression prefilter. An adaptive algorithm for implementing such a prefilter has been described in detail in reference [8]. Letting the output of this prefilter be denoted by

$$X(\omega) = P(\omega)Y(\omega) \tag{A-11}$$

then the pitch likelihood function reduces to the correlation operation

$$\ell(\tau) = \int_{0}^{\pi} C(\omega;\tau)X(\omega)X^{*}(\omega)d\omega \tag{A-12}$$

Since the function $C(\omega;\tau)$ was defined to represent the discrete periodic structure of the speech spectrum, it must be symmetric, non-negative and periodic. In the ideal case a suitable representation is

$$C(\omega;\tau) = \frac{1}{\tau} \sum_{m=-\infty}^{\infty} \delta(\omega - \frac{2\pi m}{\tau}) \tag{A-13}$$

Since speech is at most quasi-periodic, a more practical choice for $C(\omega;\tau)$ is (other choices are possible)

$$C(\omega;\tau) = \frac{1}{2}(1+\cos\omega\tau) \tag{A-14}$$

which is the DFT of the sequence

$$c_n = \frac{1}{2}\delta_n + \frac{1}{4}\delta_{n-\tau} + \frac{1}{4}\delta_{n+\tau} \tag{A-15}$$

hence the likelihood ratio can be written as

$$\ell(\tau) = \frac{1}{2}\sum_{n=1}^{N} x_n \left[x_n + \frac{1}{2}(x_{n-\tau} + x_{n+\tau})\right] \qquad \text{(A-16)}$$

Since only the second term depends on the pitch period, then the final expression for the likelihood function is

$$\ell(\tau) = \sum_{n=1}^{N} x_n \hat{x}_n(\tau) \qquad \text{(A-17)}$$

where

$$\hat{x}_n(\tau) = \frac{1}{2}(x_{n-\tau} + x_{n+\tau}) \qquad \text{(A-18)}$$

which represents the output of a comb filter tuned to pitch period τ. Of course a bank of comb filters is needed, each tuned to a different pitch period, and the one that leads to the largest correlation determines the maximum likelihood estimate of the pitch period.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

19 REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 18 ESD TR-79-181	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) 6 Design of a Robust Maximum Likelihood Pitch Estimator for Speech in Additive Noise		5. TYPE OF REPORT & PERIOD COVERED 9 Technical Note 14 TN- 6. PERFORMING ORG. REPORT NUMBER Technical Note 1979-28
7. AUTHOR(s) 10 Robert J. McAulay		8. CONTRACT OR GRANT NUMBER(s) 15 F19628-78-C-0002
9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Element No. 33401F Project No. 7280
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Systems Command, USAF Andrews AFB Washington, DC 20331	11	12. REPORT DATE 11 June 1979 13. NUMBER OF PAGES 46
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Electronic Systems Division Hanscom AFB Bedford, MA 01731	12 45	15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

narrowband digital speech
vocoders
maximum likelihood estimation
pitch estimation
additive noise

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Application of the maximum likelihood technique to a speech model that includes the effects of pitch periodicity and the spectral envelope results in a processing structure that consists of a noise suppression prefilter in cascade with a comb filter bank estimator-correlator. The comb filter interpretation leads to a pitch likelihood function which avoids the problem of anomalous errors due to the effects of windowing and formant sidelobe interaction. Pitch ambiguities are resolved using a majority logic scoring algorithm and a carefully designed pitch tracker that can adapt rapidly to gross pitch variations. A time domain implementation of the algorithm that runs in real time in conjunction with an LPC analysis/synthesis system at 2400 bps is described.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

207650